IENG3304 – DATA MANAGEMENT & ANALYTICS

Report for LAB ASSIGNEMENT 2:

WRITTEN BY: ABILASH SURENDRAN

BANNER ID: B00891410

Table Of content:

LIST OF FIGURES:

Introduction:

The red wine industry has recently expanded as social drinking becomes increasingly popular. In such circumstances, wine ratings rise. Several variables contribute to the wine's higher ranking. Winery companies utilise these ratings to advertise and market their products. This database contains information about red Spanish wine types. The dataset shows how the quality of the dataset is affected by different popularity and description variables. The purpose of this lab assignment is to comprehend the relationship between numerous elements that influence the overall rating of the wine, as well as to create a regression model that includes both rating and price as dependent variables in order to get insight into how each aspect influences them.

Body:

Before purchasing wine, a wine aficionado must consider several variables. The winery, wine, type, body, acidity, number of reviews and ratings, as well as the region and country where the winery is located, are all aspects to consider. For an end consumer, however, not all these factors are essential. For example, information about the vineyard adds little value to the end client, yet the wine's "BRAND NAME" can attract multiple buyers if it has a large market share. Ratings, the number of reviews, the acidity level, and the body all play a role in assisting the customer. Let's take a closer look at each of these variables:

Winery: A winery is just a wine producing facility where wine is made and stored. Teso La Monja and Artadi are two winery companies in "Espana" that are less expensive than the rest but nevertheless have a market grade of 4.9. This information is derived from the dataset.

Wine: The “BRAND NAME” of several wines are listed here.

Rating: Ratings are simply the overall consensus on the wine. This is the average of all the feedback received. It ranges from 4.2 to a maximum of upto 4.9. Also, there are other factors price and year that also enhances the overall rating of the wine.

HOW DOES THE NUMBER OF REVIEWS INFLUENCE THE RATING OF WINE?

The number of reviews is used to predict the wine's rating. The more the number of reviews is, the better is the rating obtained. The number of reviews, on the other hand, might be influenced by the price. If the wine is of good quality and reasonably priced, it has a decent chance of receiving the highest number of ratings.

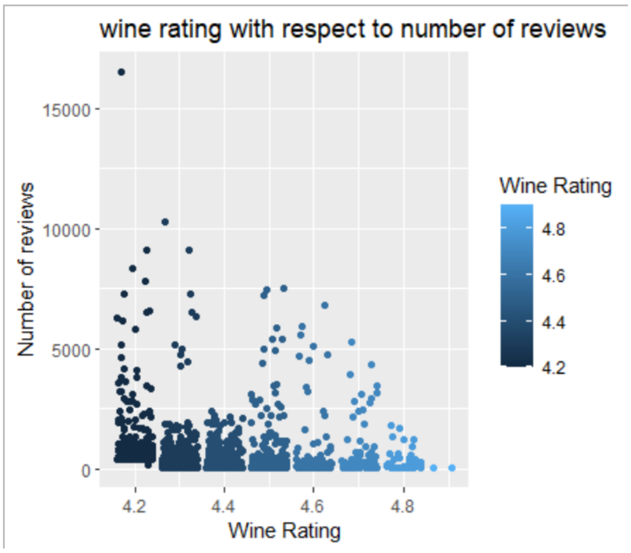


Fig 1: Impact of wine rating with respect to reviews

For a better understanding of how the number of reviews influences the rating, look at the graph above. The number of reviews for the 4.2 rating had clearly increased to 16000. When there are a lot of evaluations, however, the average is used, lowering the overall rating of multiple wines to around 4.2. In comparison to other wines, wines with a rating of 4.9 have comparatively few reviews. As a result, the wine with the most reviews is more promising, whilst the wine with the fewest reviews is not.

WHAT IS THE RELATIONSHIP BETWEEN PRICE OF WINE AND RATING OF THE WINE?

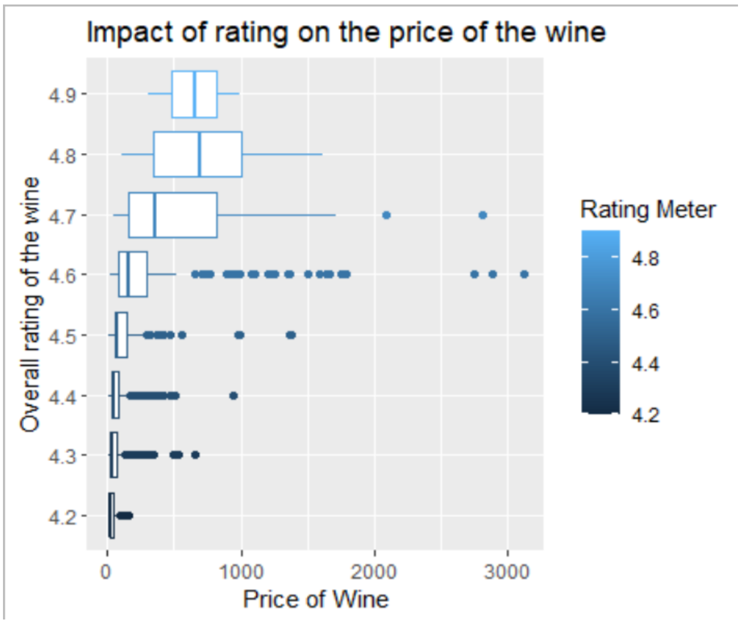


Fig 2: Impact of rating on the price of the wine

The box plot is a graphic representation of the link between rating and price. The price of different wines is spread throughout a scale of 0 to 3000 for each rating group. The outliers are clearly found in the ratings range of 4.7 and lower. This shows that there are various other wines with similar ratings but different price points. Each box plot's middle line represents the pricing range's median.

Body:

Body refers to the wine’s richness (or) quality, which becomes apparent as soon as it is drunk for the first time. Several chemical components such as alcohol, sulphides, chlorine and others, could boost the richness.

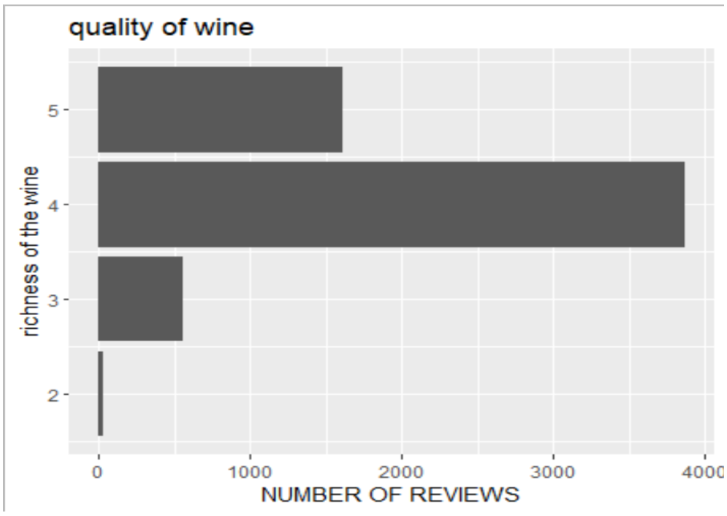


Fig 3: Quality of wine

The given dataset's body value varies from 2 to 5, with a minimum of 2. The chemical makeup of the wine with a body value of 5 as the maximum richness value is of top quality. It can be deduced that luxury wines are more expensive than the others. As a result, it is consumed by a comparatively small number of people. As a result, a consumer would be interested in learning more about the wine he or she is about to enjoy.

TYPE:

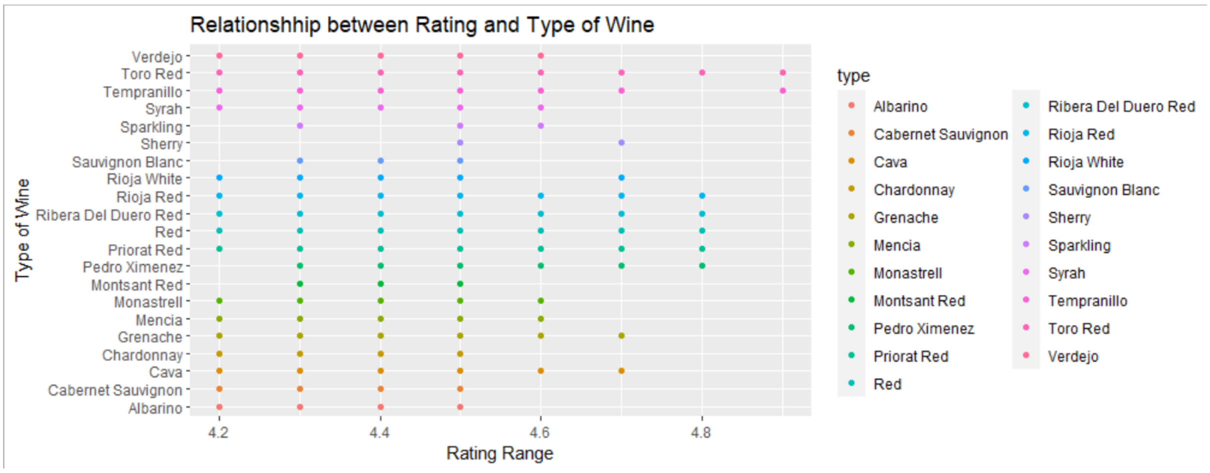


Fig 4: relationship between the type of wines and rating

This chart depicts the relationship between the type of wine and rating. There are totally 21 different type of wines which is spread across a rating range from 4.2 to 4.9. Since this dataset is spread across a period of time (from 1925 to 2022), we can notice that a same type of wine as different ratings.

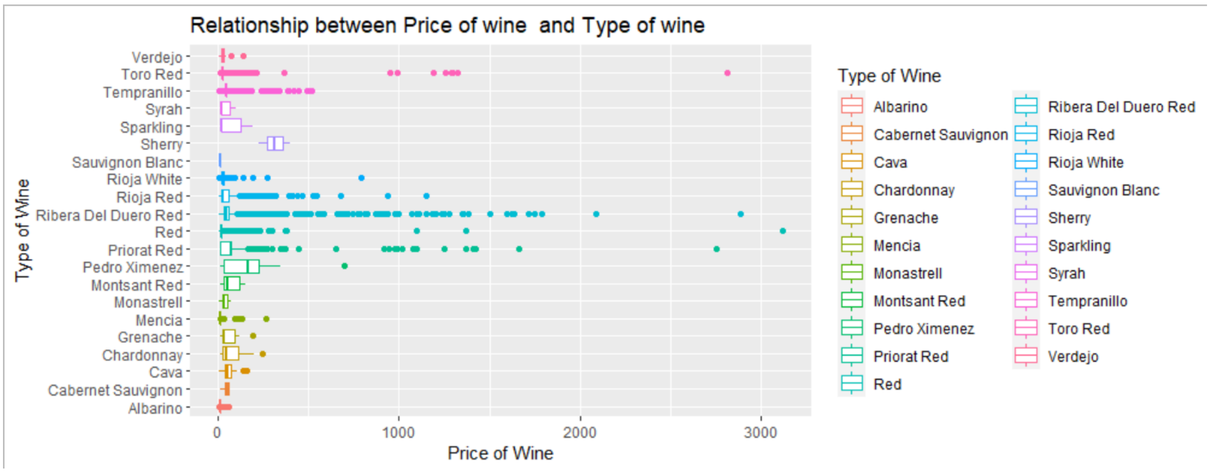


Fig 5: Relationship between price of wine and type of wine

This chart is used to understand the relationship with the cost of wine and different type of the wine. It can be noticed that there are several outliers in the plot. It means that the price of the wines have fluctuated over the period of time.

Regression model:

Regression model is generally used to predict the relationship between the dependent variable and the independent variables.

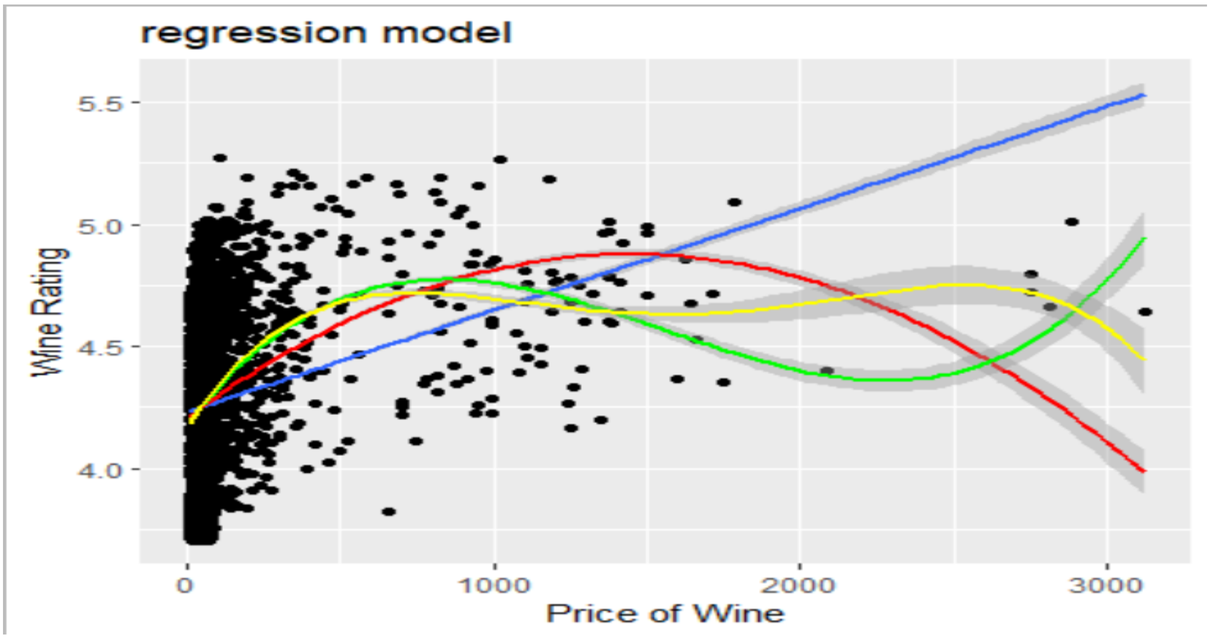


Fig 6: regression model

Here four different regression models are created whose degrees are 1,2,3, and 4. It is clear from the graph that the lines of order 2,3 and 4 that is line in red, green and yellow are clearly overfitting the dataset. It also increases the complexity of the model and thereby makes it difficult to interpret values from the graph.

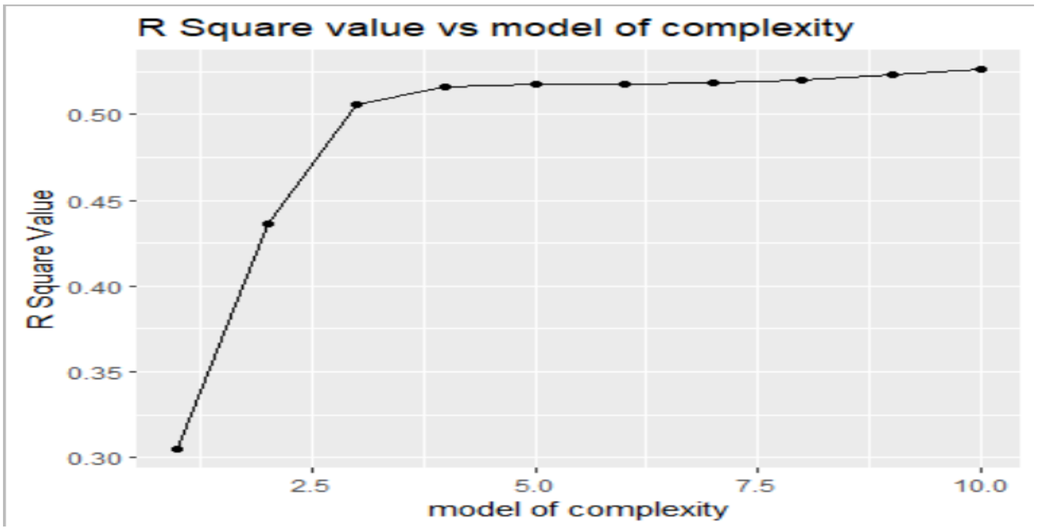


Fig 7: R square value with linear regression model vs model of complexity

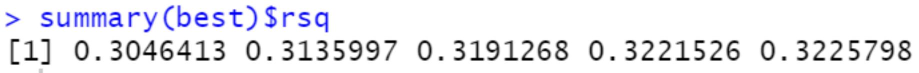
The corresponding R square values are plotted for 10 different models whose complexity increases with increase in order. It is clear from the graph that model 3 whose R square value is 51% is comparatively the model, as any model with increase in complexity after model 3 leads to overfitting.

For identifying the optimum method to enhance the R- square value, the best subset function was used.

1. Firstly, the “RATING” of the wine is considered as a dependent variable which acts as the function of year, number of reviews, acidity, body and price. Year is considered as a variable to identify the overall change in the rating of a particular wine over a period of time. Moreover, the acidity and quality of the wine (body) could also change as time progresses. Finally, the number of reviews had also fluctuated throughout. Both linear and logarithmic models were created to obtain the R – squared value. Five different models were created and compared to identify the best model which produces improved R- squared value.

For best model:

Ratings was maintained as a function of year, price, acidity, body, number of reviews :



```
> summary(best)$rsq
[1] 0.3046413 0.3135997 0.3191268 0.3221526 0.3225798
```

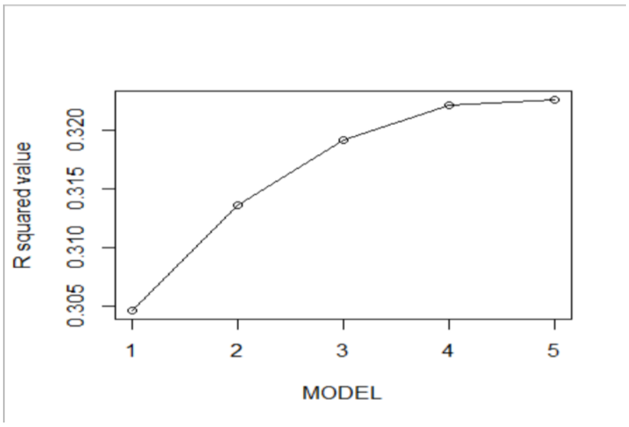
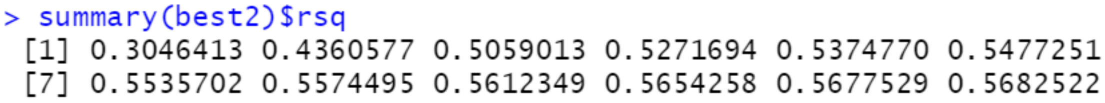


FIG 8:Model 1: Improved R-Squared Value For Dependent Variable Rating

For best2 model:

Here rating is set up with every variable to its polynomial function



```
> summary(best2)$rsq
 [1] 0.3046413 0.4360577 0.5059013 0.5271694 0.5374770 0.5477251
 [7] 0.5535702 0.5574495 0.5612349 0.5654258 0.5677529 0.5682522
```

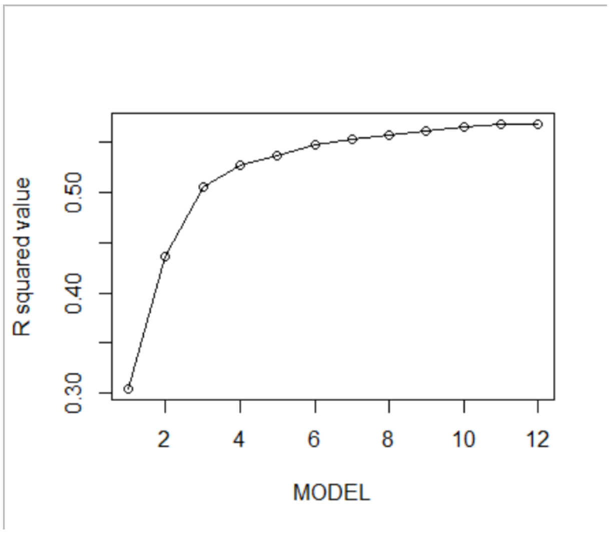
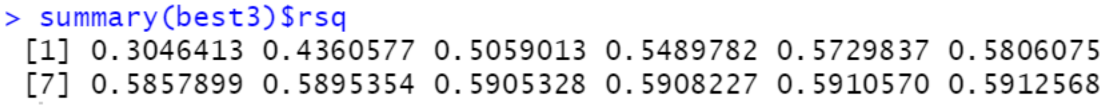


Fig 9: Model 2: Improved R-Squared Value For Dependent Variable Rating

best3 model:

Here rating is a function of year, acidity and also to the polynomial function of number of reviews, price, body.



```
> summary(best3)$rsq
 [1] 0.3046413 0.4360577 0.5059013 0.5489782 0.5729837 0.5806075
 [7] 0.5857899 0.5895354 0.5905328 0.5908227 0.5910570 0.5912568
```

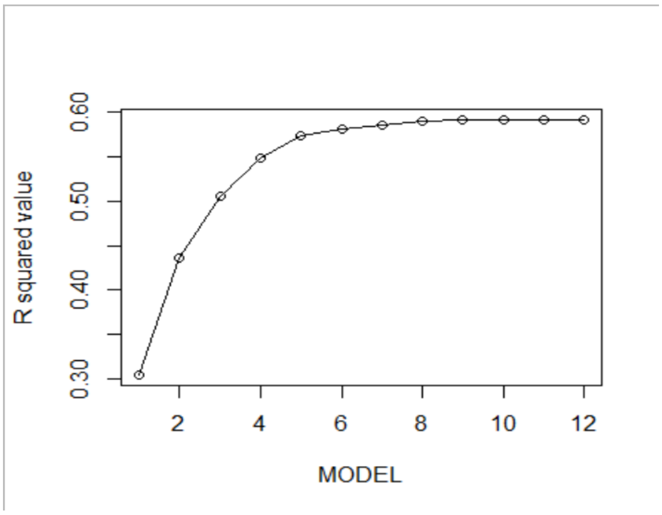
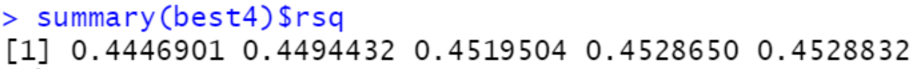


FIG 10: Model 3:Improved R-Squared Value For Dependent Variable Rating

best4 model:

Here rating is a function of year, number of reviews, body, acidity, and on top of it a logarithmic function of price.



```
> summary(best4)$rsq
[1] 0.4446901 0.4494432 0.4519504 0.4528650 0.4528832
```

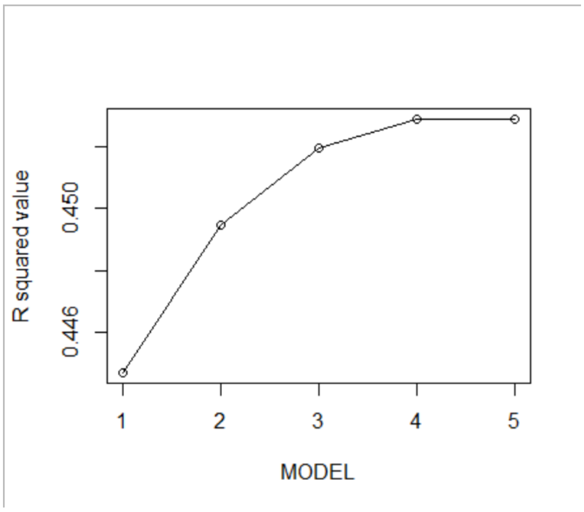
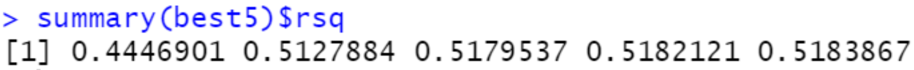


Fig 11:Model 4: Improved R-Squared Value For Dependent Variable Rating

Best5 model:

Here rating is a function of year, body, acidity, and on top of it a logarithmic function of price and number of reviews.



```
> summary(best5)$rsq
[1] 0.4446901 0.5127884 0.5179537 0.5182121 0.5183867
```

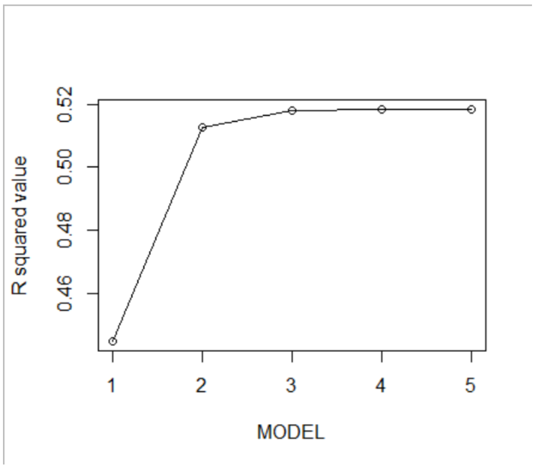


Fig 12:Model 5: Improved R-Squared Value For Dependent Variable Rating

Therefore to enhance the R square value with the variable RATING as the dependent variable, 5 models were created Out of which best3 is the best in the linear model and best 5 is the best in logrithmic model.

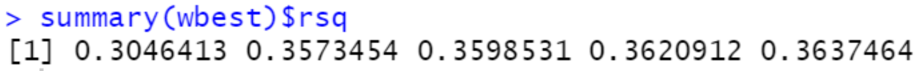
In best3 the R squared value started from a minimum of 30% and went upto a maximum of 59% . It had created upto 12 models. The 5th and 6th model of best3 as a R square value of 57% and 58% which is considered to be the best, as the R square values of the remaining models just had a very small increase which can be considered as negligible. Also selecting any model after 6th will result in overfitting.

On the other hand, in the best5 which uses the logrithmic function the R squared values started at 44% and had consistently risen upto 51%. It had created about 5 models , out of which the second model is considered to be the best. Its R square value is 51%. All models after 2nd have a minor increase in its R square value which can be neglected.

2. Secondly, the "price" of the wine is considered as a dependent variable which acts as the function of year, number of reviews, acidity, body and rating. Both linear and logarithmic models were created to obtain the R – squared value. Five different models were created and compared to identify the best model which produces improved R- squared value.

MODEL 1:

Price was maintained as a function of year, rating, acidity, body, number of reviews :



```
> summary(wbest)$rsq
[1] 0.3046413 0.3573454 0.3598531 0.3620912 0.3637464
```

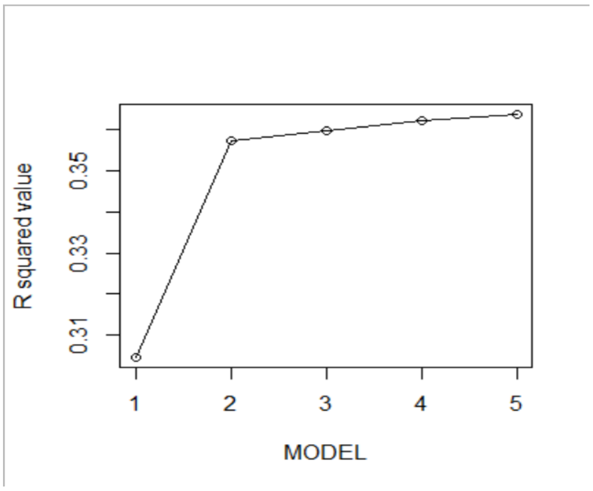
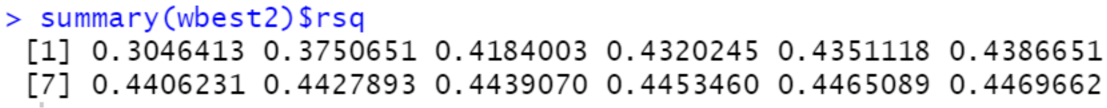


Fig 13:Model 1: Improved R-Squared Value For Dependent Variable Price

MODEL 2:

Here price is set up with every variable to its polynomial function



```
> summary(wbest2)$rsq
 [1] 0.3046413 0.3750651 0.4184003 0.4320245 0.4351118 0.4386651
 [7] 0.4406231 0.4427893 0.4439070 0.4453460 0.4465089 0.4469662
```

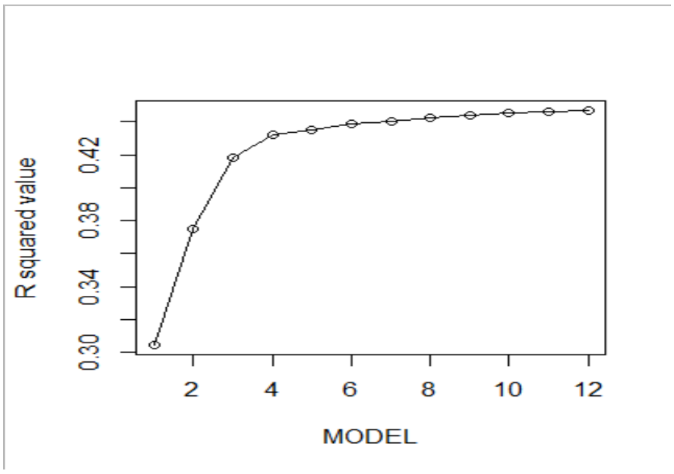
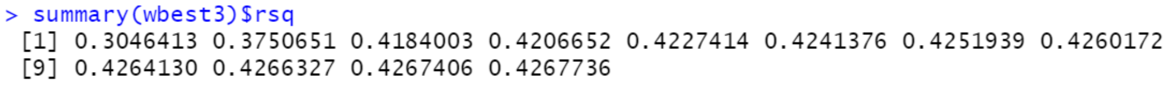


Fig 14:Model 2: Improved R-Squared Value For Dependent Variable Price

MODEL 3:

Here rating is a function of year, acidity and also to the polynomial function of number of reviews, rating, body.



```
> summary(wbest3)$rsq
 [1] 0.3046413 0.3750651 0.4184003 0.4206652 0.4227414 0.4241376 0.4251939 0.4260172
 [9] 0.4264130 0.4266327 0.4267406 0.4267736
```

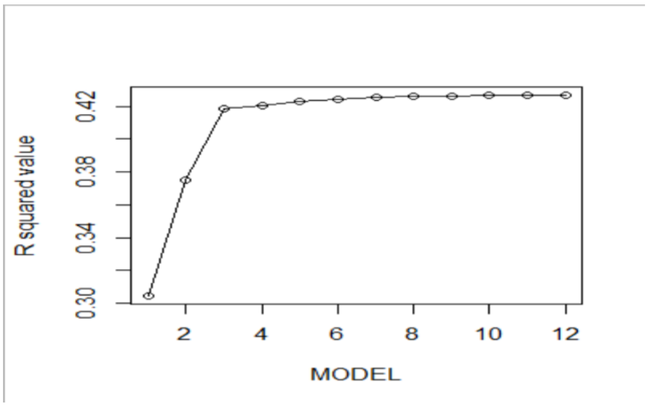
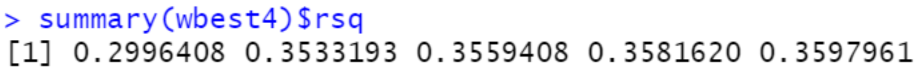


Fig 15:Model 3: Improved R-Squared Value For Dependent Variable Price

MODEL 4:

Here price is a function of year, number of reviews, body, acidity, and on top of it a logarithmic function of rating.



```
> summary(wbest4)$rsq
[1] 0.2996408 0.3533193 0.3559408 0.3581620 0.3597961
```

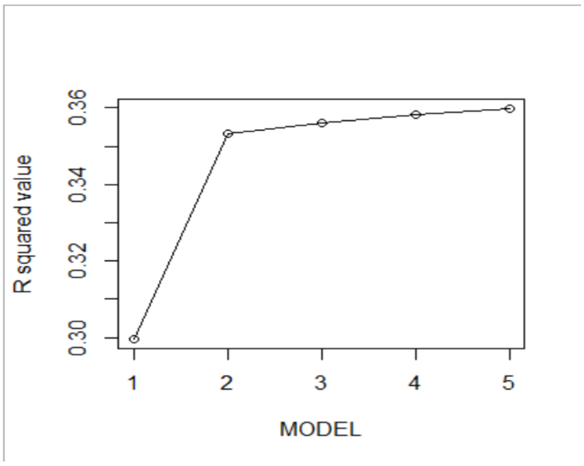
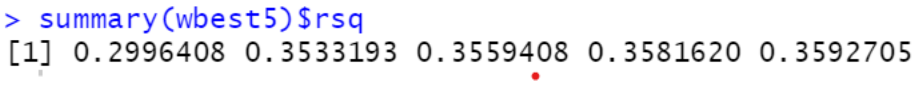


Fig 16:Model 4: Improved R-Squared Value For Dependent Variable Price

MODEL 5:

Here price is a function of year, body, acidity, and on top of it a logarithmic function of rating and number of reviews.



```
> summary(wbest5)$rsq
[1] 0.2996408 0.3533193 0.3559408 0.3581620 0.3592705
```

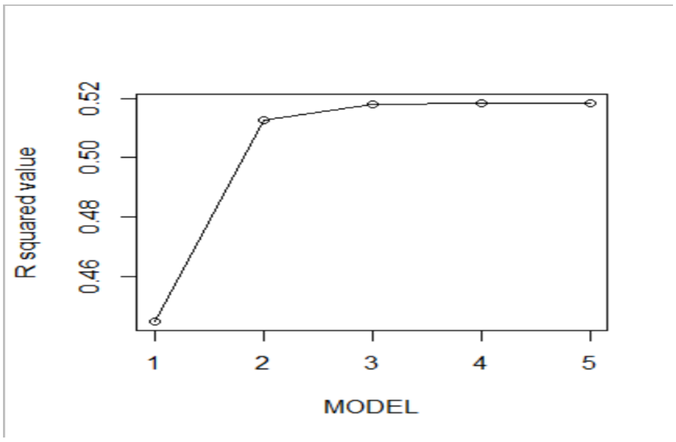


Fig 17:Model 5: Improved R-Squared Value For Dependent Variable Price

Here model2 is the best in the linear model . Moreover, there was not much difference with model 4 and model 5. Therefore any of the two logrithmic model can be selected. In model 2 the R squared value started from a minimum of 30% and went upto a maximum of 44% . It had created upto 12 models. The 4th model of MODEL 2 as a R square value of 43% which is considered to be the best, as the R square values of the remaining models just had a very small increase which can be considered as negligible. Therefore while considering "PRICE" as the dependent variable, model 2 can be used to obtain the best R Square values.

Conclusion:

In this lab assignment a wine dataset is taken to understand the regression model. Both price and rating were taken as dependent variables and the corresponding linear regression model were created. To enhance the R- square values, the best subset function was used. In total, 10 different models were created, and the best model for price and rating is selected and discussed in the report.

Reference:

Curated dataset from Brightspace: https://dal.brightspace.com/d2l/le/content/221958/viewContent/3012797/View

R- CODE:

```
#assignment 2

library(tidyverse)

library(dplyr)

library(scales)

library(ggplot2)
library(leaps)

#setting up working directory
setwd("C:/CANADA/industrial engineering/summer term/data analytics/dataset")
#importing the dataset
dataset= read.csv("spanish_wine.csv")
view(dataset)
dim(dataset)
str(dataset)
summary(dataset)

colSums(is.na(dataset))
cleandata <- dataset[complete.cases(dataset),]
dim(cleandata)
cleandata$year <- as.integer(cleandata$year)

view(cleandata)
#cleaning dataset
New_wine = drop_na(cleandata)
view(New_wine)
str(New_wine)
dim(New_wine)

#models to increase the r2value
wine= dplyr::select(New_wine, year,rating, num_reviews, body, acidity, price)
best = regsubsets(rating~., data=wine)
summary(best)
summary(best)$rsq
plot(summary(best)$rsq, type = "o",xlab = "MODEL",ylab = "R squared value")
```

```
#model2
best2 = regsubsets(rating~poly(year,3)+poly(num_reviews,3)+poly(body,3)+
                  poly(acidity,2)+poly(price,3),
                data= wine,nvmax = 12)


summary(best2)
summary(best2)$rsq
plot(summary(best2)$rsq, type = "o",xlab = "MODEL",ylab = "R squared value")


#model3
best3 = regsubsets(rating~year+poly(num_reviews,4)+poly(body,3)
+acidity+poly(price,3),data= wine,nvmax = 12)
summary(best3)
summary(best3)$rsq
plot(summary(best3)$rsq, type = "o",xlab = "MODEL",ylab = "R squared value")


#model4
best4 = regsubsets(rating~year+num_reviews+body+acidity+log(price),data= wine,nvmax
= 12)
summary(best4)
summary(best4)$rsq
plot(summary(best4)$rsq, type = "o",xlab = "MODEL",ylab = "R squared value")


#model5
best5 = regsubsets(rating~year+log(num_reviews)+body+acidity+log(price),data=
wine,nvmax = 12)
summary(best5)
summary(best5)$rsq
plot(summary(best5)$rsq, type = "o",xlab = "MODEL",ylab = "R squared value")
```

```
ggplot(data=New_wine, aes(y=rating,x=price))+
  geom_point(position = position_jitter(width = 1, height = .5))+
  geom_smooth(method="lm")+
  geom_smooth(method="lm",formula= y~poly(x,2),colour= "red")+
  geom_smooth(method="lm",formula= y~poly(x,3),colour= "green")+
  geom_smooth(method="lm",formula= y~poly(x,4),colour= "yellow")+
  xlab("Price of Wine")+
  ylab("Wine Rating")+
  ggtitle("regression model")

rval= seq(1,10)
rvalx= seq(1,10)
for (i in 1:10) {
  model= lm(data=New_wine,rating~poly(price,i))
  rval[i]= summary(model)$r.squared
}
view(rval)

modelR2val= data.frame(rvalx,rval)
ggplot(data=modelR2val)+
  geom_point(aes(x=rvalx,y=rval))+
  geom_path(x=rvalx,y=rval)+
  xlab("model of complexity")+
  ylab("R Square Value")+
  ggtitle("R Square value vs model of complexity")

#regression model with dependent variable to be price

wines= dplyr::select(New_wine, year,rating, num_reviews, body, acidity, price)
wbest = regsubsets(price~., data=wines)
```

```
summary(wbest)
summary(wbest)$rsq
plot(summary(wbest)$rsq, type = "o",xlab = "MODEL",ylab = "R squared value")

#model2
wbest2 = regsubsets(price~poly(year,3)+poly(num_reviews,3)+poly(body,3)+
                  poly(acidity,2)+poly(rating,3),
                data= wines,nvmax = 12)
summary(wbest2)
summary(wbest2)$rsq
plot(summary(wbest2)$rsq, type = "o",xlab = "MODEL",ylab = "R squared value")

#model3
wbest3 = regsubsets(price~year+poly(num_reviews,4)+poly(body,3)
+acidity+poly(rating,3),data= wines,nvmax = 12)
summary(wbest3)
summary(wbest3)$rsq
plot(summary(wbest3)$rsq, type = "o",xlab = "MODEL",ylab = "R squared value")

#model4
wbest4 = regsubsets(price~year+num_reviews+body+acidity+log(rating),data=
wines,nvmax = 12)
summary(wbest4)
summary(wbest4)$rsq
plot(summary(wbest4)$rsq, type = "o",xlab = "MODEL",ylab = "R squared value")

#model5
wbest5 = regsubsets(price~year+log(num_reviews)+body+acidity+log(rating),data=
wines,nvmax = 12)
summary(wbest5)
summary(wbest5)$rsq
```

```
plot(summary(best5)$rsq, type = "o",xlab = "MODEL",ylab = "R squared value")

#plotting the graph for various variables
#reviews vs ratings
ggplot(data=New_wine)+
  geom_bar(mapping = aes(y=body))+
  ylab("richness of the wine")+
  xlab("NUMBER OF REVIEWS")+
  ggtitle("quality of wine")

ggplot(data = New_wine)+
  geom_boxplot(mapping = aes(y = factor(rating), x = price, colour = rating))+
  ylab("Overall rating of the wine")+
  xlab("Price of Wine")+
  labs(color = "Rating Meter")+
  ggtitle("Impact of rating on the price of the wine")

ggplot(data = New_wine)+
  geom_boxplot(mapping = aes (y = factor(type), x = price, colour = type))+
  ylab("Type of Wine")+
  xlab("Price of Wine")+
  labs(color = "Type of Wine")+
ggtitle("Relationship between Price of wine  and Type of wine")

ggplot(data = New_wine)+
  geom_point(mapping = aes (x = rating, y = type, color = type))+
  ylab("Type of Wine")+
  xlab("Rating Range")+
  ggtitle("Relationshhip between Rating and Type of Wine")
```

```
ggplot(data = New_wine)+
  geom_point(mapping=aes(x=rating, y=num_reviews, color = rating), position= "jitter")+
  ylab("Number of reviews")+
  xlab("Wine Rating")+
  labs(color = "Wine Rating")+
  ggtitle("wine rating with respect to number of reviews")
```